Optimizing Overlay-based Virtual Networking Through Optimistic Interrupts and Cut-through Forwarding

Zheng Cui[†] — [†]University of New Mexico
Lei Xia[‡] — [‡]Northwestern University
Patrick Bridges[†] — [†]University of New Mexico
Peter Dinda[‡] — [‡]Northwestern University
Jack Lange[*] — [*]University of Pittsburgh

http://v3vee.org

Overview

- **Motivation:**
  - Overlay-based virtual networks
  - Bandwidth and latency limitations

- **Core issues:**
  - Delayed and/or excessive virtual interrupts
  - Copies between guest and host data buffers

- **Key optimizations:**
  - Optimistic timer-free virtual interrupt injection
  - Zero-copy, cut-through data forwarding

- **Performance evaluation on 10Gbps Ethernet:**
  - Latency: reduced by 50%
  - Throughput: increased by > 30%
  - Near-native application performance

Motivations

- **Virtual overlays are important for cloud systems**
  - Easy deployment/management
  - Location/Hardware independence
- **Evaluated performance of VNET/P overlay**
- **Performance limitations on faster networks (e.g.,10Gbps Ethernet):**
  - Latency: 3 times higher than native
  - Throughput: ~60% of native
  - Large latency variation
  - 30-40% HPCC application benchmark slowdown

Linux Host + Palacios VMM + VNET/P

- **Palacios VMM**
  - OS-independent embeddable virtual machine monitor
  - Open source
  - Host OS: Linux, Kitten, Minix ...
- **VNET/P**
  - Layer 2 virtual overlay network for user's virtual machines
  - Virtual NIC for each guest OS
  - VNET core
  - VNET bridge

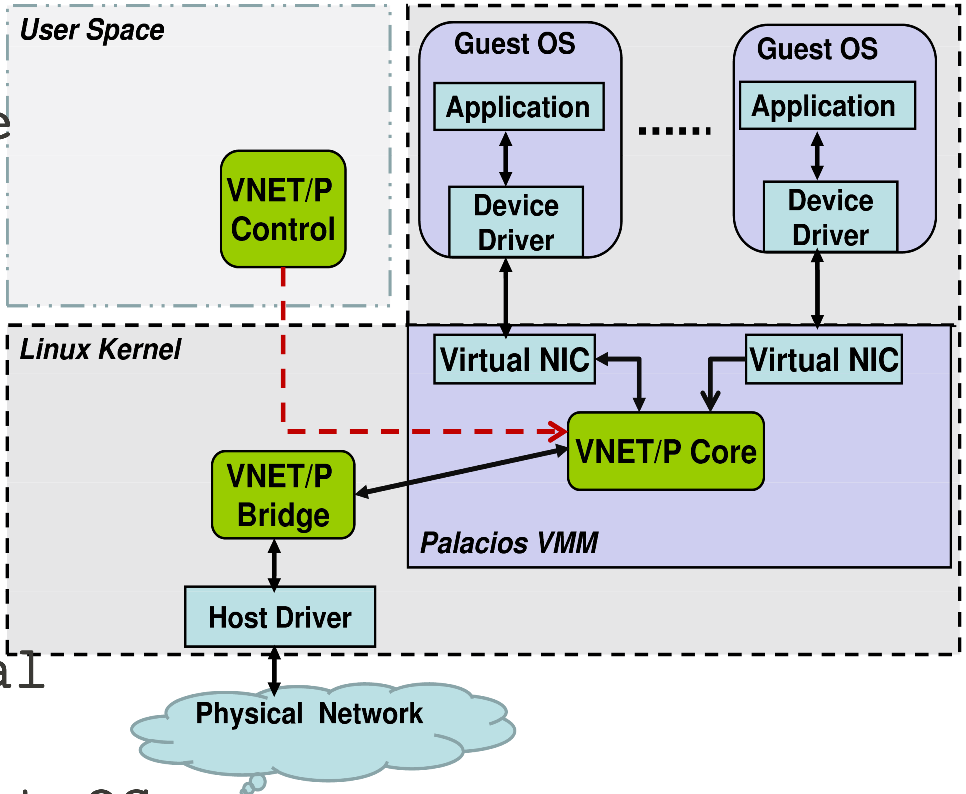


Fig. 1. VNET/P architecture.

Performance Challenges

- **Delayed virtual interrupts**
- **Excessive virtual interrupts**
- **High-resolution timer noise**

Delayed virtual interrupts

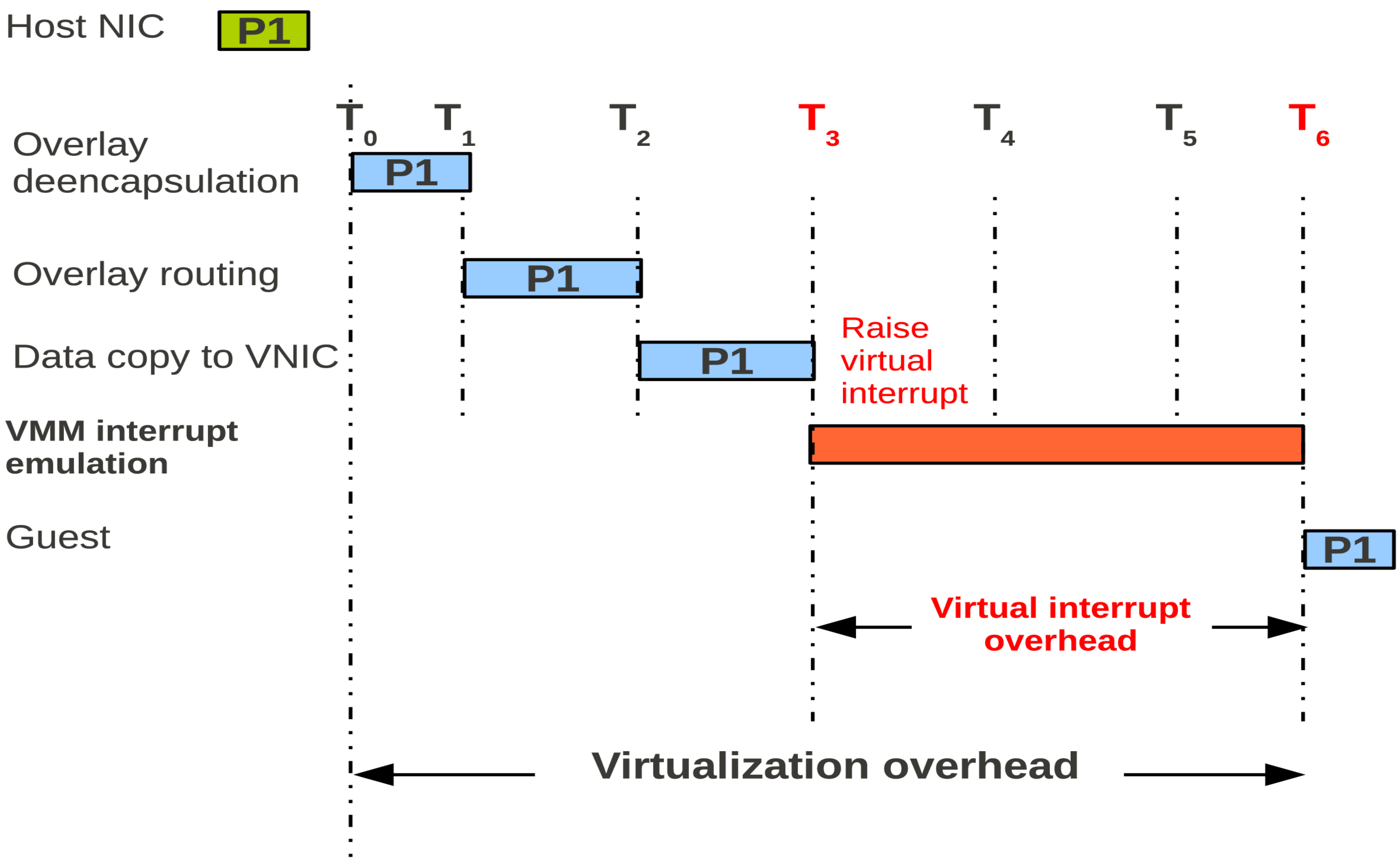


Fig.2 Packet Processing Time Line

Performance Challenges

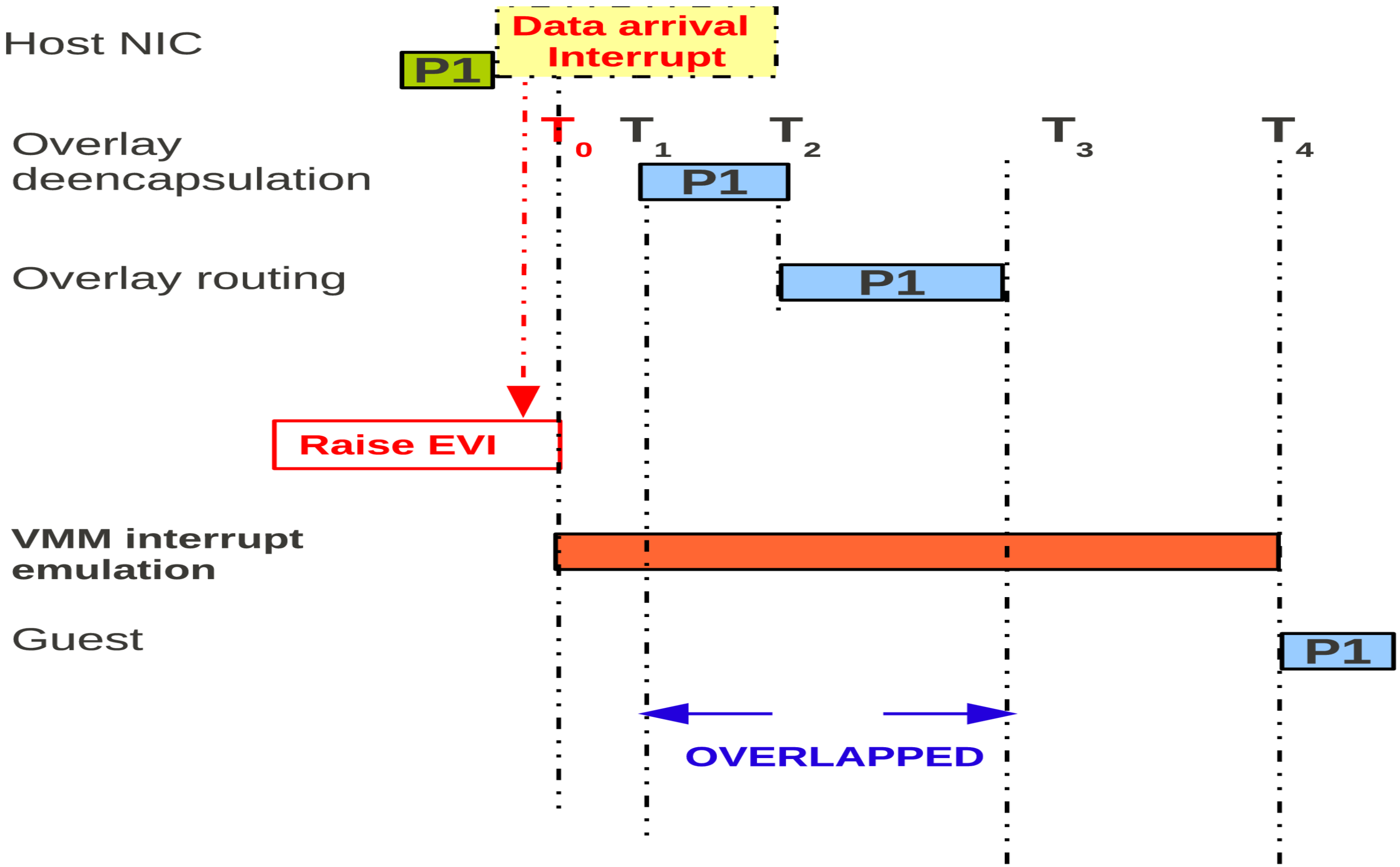
- Delayed virtual interrupts
- **Excessive virtual interrupts**
- **High-resolution timer noise**

Optimization Overview

- **Optimizations:**
  - Optimistic Interrupts
  - Zero-copy cut-through data forwarding
- **Leverage a low-noise host OS**
- **Assumption:**
  - One-to-one binding of host and virtual NIC receive queues
  - Capability provided by modern NICs

Optimization# 1:
Optimistic Interrupts

- Early Virtual Interrupt (EVI) delivery
- End of Coalescing (EoC) notification



Early Virtual Interrupt（EVI）delivery

Three scenarios:

1. **Virtual interrupts disabled:**
    - Discard by VMM
    - Implicitly coalesced with a later interrupt

2. **Guest handler runs prior to packet availability:**
    - Ignores by guest
    - Wasting guest OS CPU

3. **Guest handler runs after packet availability:**
    - Not early enough
    - Latency increases
    - Extreme scenario: unoptimized VNET/P

End of Coalescing (EoC) notification

- **Problem:**
  - EVI delivery may fail
  - Guest's processing may out-pace overlay's processing
- **Solution:** Raise interrupt when host receive queue empty
  - Host device driver sends EoC to overlay
  - Overlay injection based on:
    - Previous EVI success/failure
    - Shape of the traffic since last EVI
- **Impact:**
  - Bound packet latency without high-resolution timers
  - Additional benefit: avoid excessive virtual interrupts

Optimization#2: **Zero-copy cut-through data forwarding**

- **Goals:**
  - Increase interrupt efficiency
  - Synchronize guest/overlay processing
- **Approach:**
  Directly forward incoming/outgoing packets between virtual NICs and host NICs
- **Mechanism:** DMA from host NIC to virtual NIC

Noise isolation to reduce performance variation

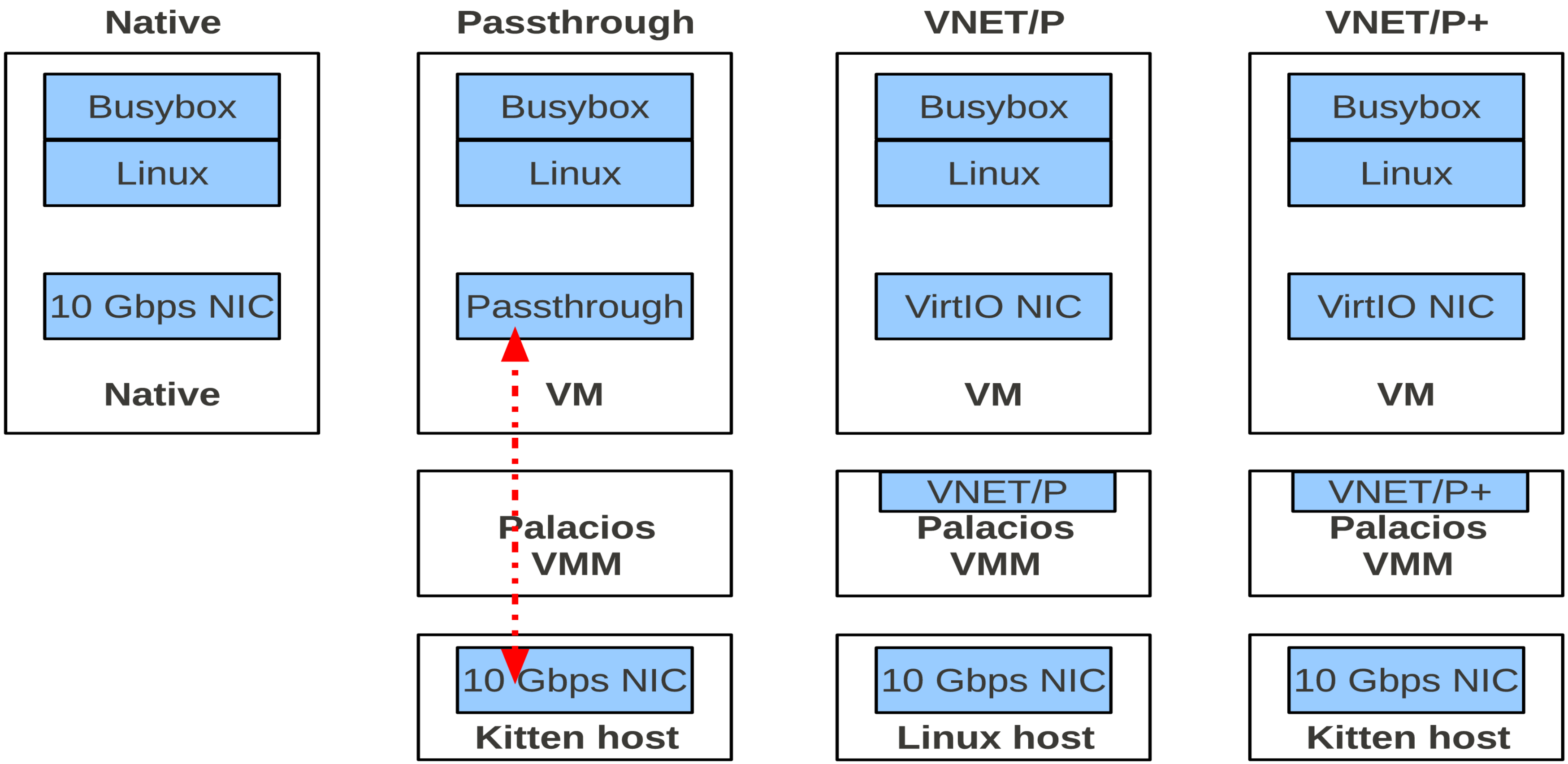
- **Approach:** Reduce host OS timer noise

- **Impacts:**
  - Reduces network performance variability
  - Increases the effectiveness of optimistic interrupts

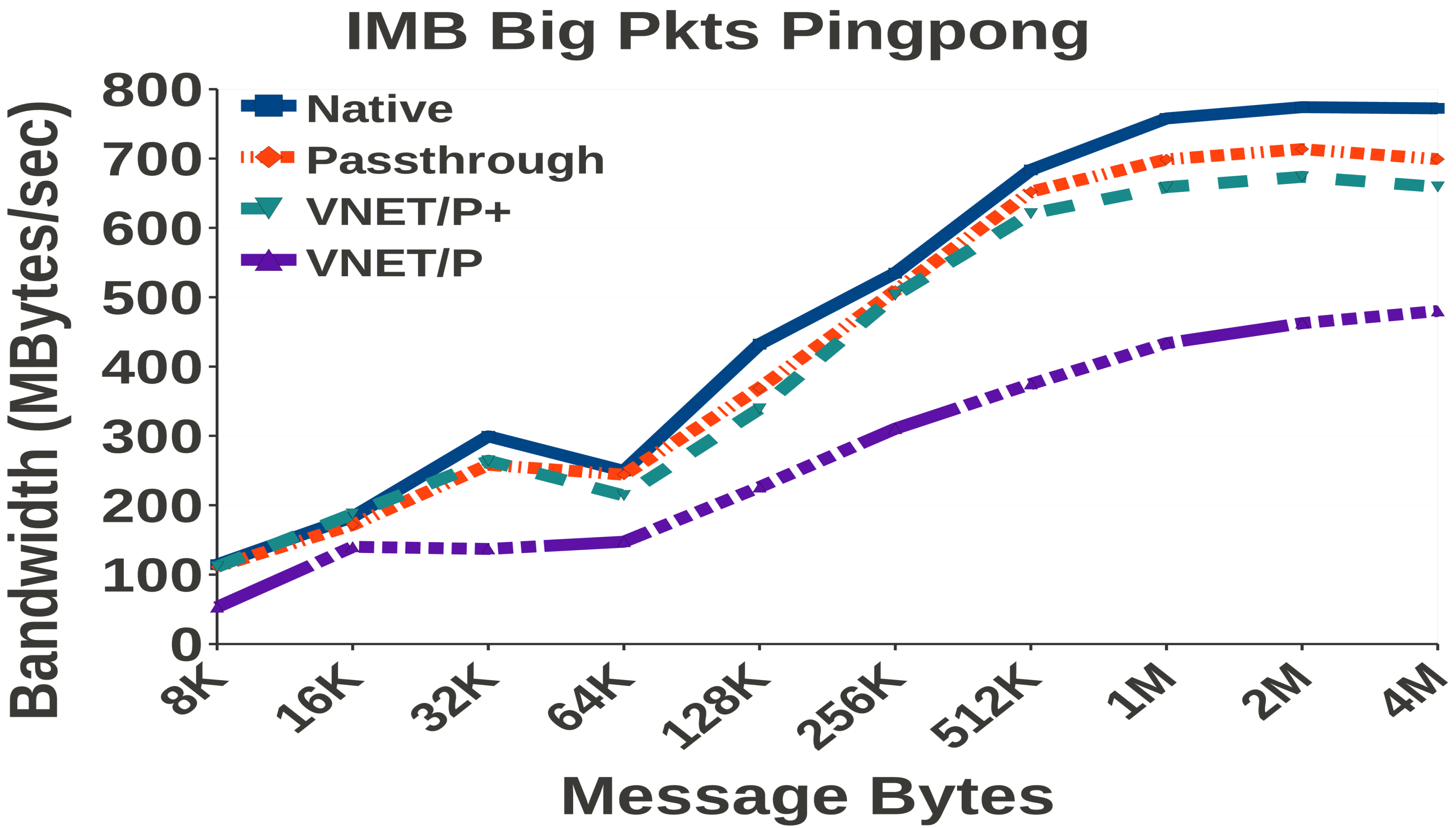
- **Implementation:** Sandia Kitten lightweight kernel

Testbed

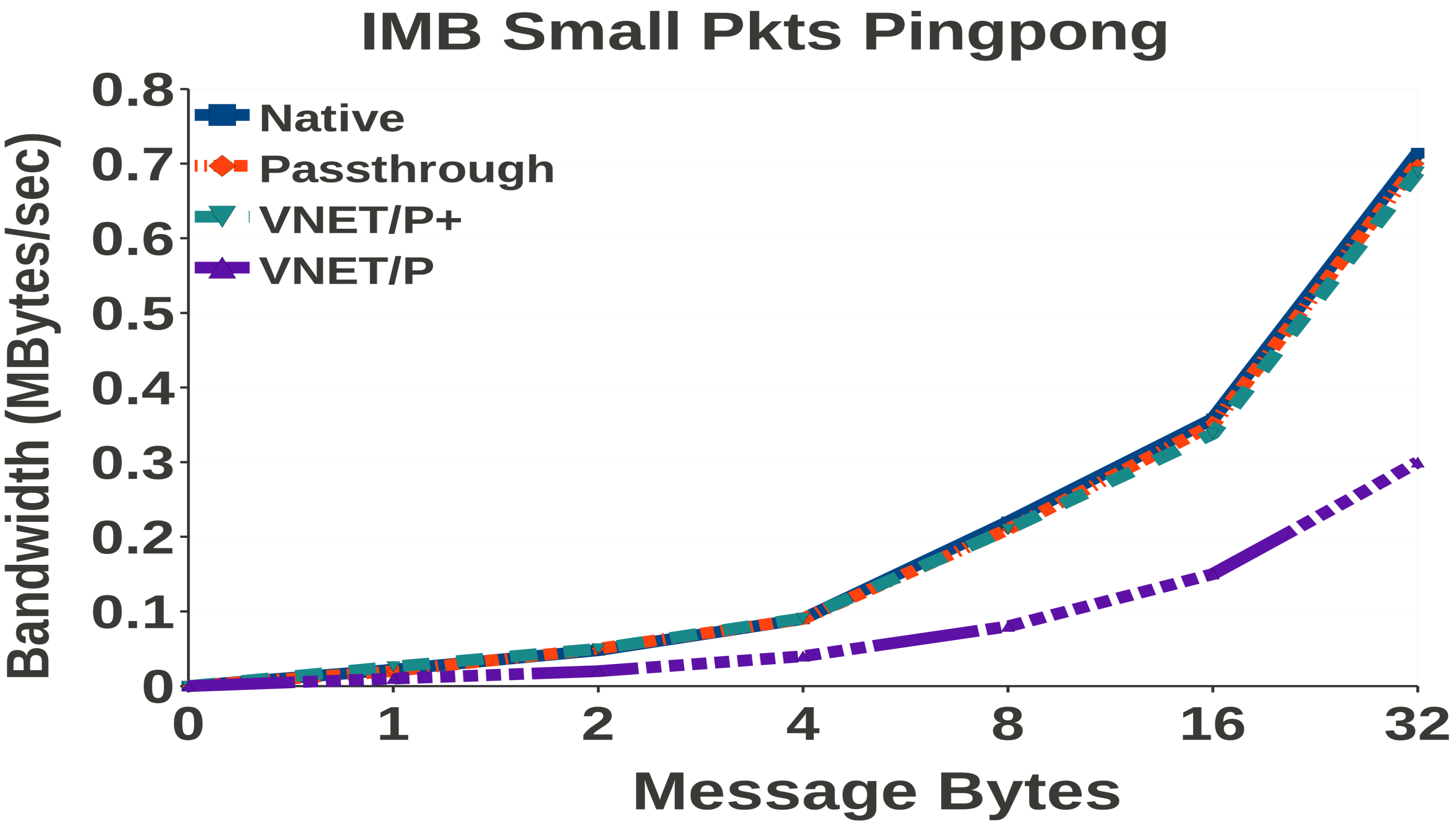
- **6-node cluster:** 8-core AMD Opteron CPU + 32GB RAM + NetEffect NE020 10Gbps Ethernet NIC
- **Configuration:**



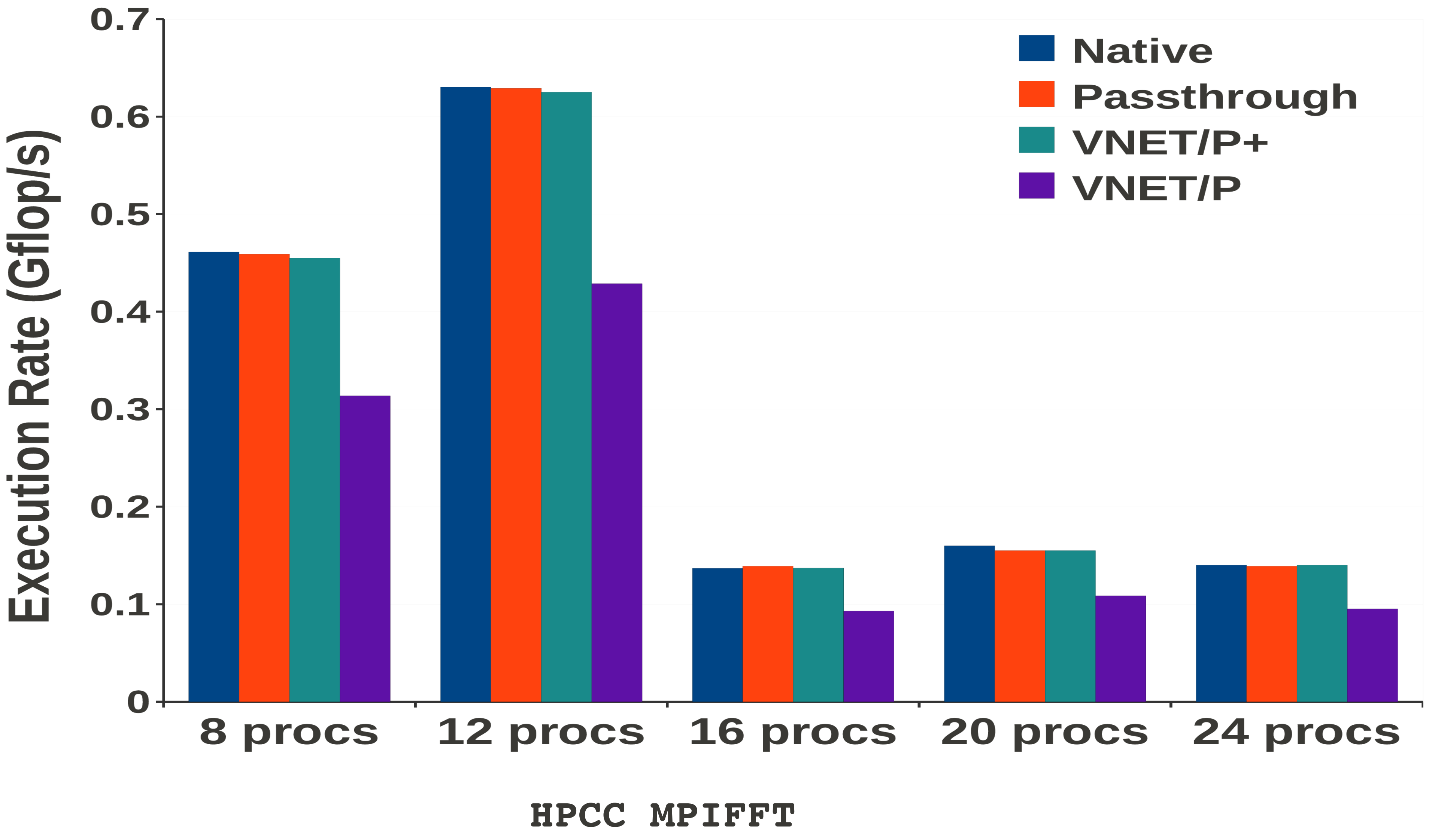
VNET/P+: Near-native MPI P2P Bandwidth



VNET/P+: Near-native MPI P2P Latency



VNET/P+: Native HPCC MPI Application Performance



VNET/P+: Near-native NAS Application Performance

Mop/s	Native	Passthrough	VNET/P	**VNET/P+**	$\frac{VNET/P+}{Native}$ (%)
ep.B.8	102.18	102.17	102.12	**102.12**	**99.9%**
ep.B.16	208	207.96	206.25	**207.93**	**99.9%**
ep.C.8	103.13	102.76	102.14	**103.08**	**99.9%**
ep.C.16	206.22	205.39	203.98	**204.98**	**99.4%**
mg.B.8	5110.29	4662.53	3796.03	**4643.67**	**90.9%**
mg.B.16	9137.26	8384.93	7405	**8262.08**	**90.4%**
cg.B.8	2096.64	1824.05	1806.57	**1811.14**	**86.4%**
cg.B.16	592.08	592.05	554.91	**592.07**	**99.9%**
ft.B.8	2055.435	2055.4	1562.1	**2055.3**	**99.9%**
ft.B.16	1432.3	1432.2	1228.39	**1432.18**	**99.9%**
is.B.8	59.15	59.14	59.04	**59.13**	**99.9%**
is.B.16	23.09	23.05	23	**23.04**	**99.8%**
is.C.8	132.08	132	131.87	**132.04**	**99.9%**
is.C.16	77.77	77.12	76.94	**77.1**	**99.9%**
lu.B.8	7173.65	6730.23	6021.78	**6837.06**	**95.3%**
lu.B.16	12981.86	11630.65	9643.21	**12198.65**	**94%**
sp.B.9	2634.53	2634.5	2421.98	**2634.5**	**99.9%**
sp.B.16	3010.71	3009.5	2916.81	**2954.16**	**98.1%**
bt.B.9	5229.01	4750.4	4076.52	**4798.63**	**91.8%**
bt.B.16	6315.11	6314.1	6105.11	**6242.83**	**99%**

Conclusion

- **Virtual Overlay networks can achieve near-native MPI application performance**

- **Challenges in virtual overlay networks:**
  - Delayed virtual interrupts
  - Excessive virtual interrupts
  - High-resolution timer noise

- **Optimization approaches:**
  - Optimistic interrupts
  - Cut-through forwarding

- **Optimization efficiency:**
  - Latency: reduced by 50%
  - Throughput: increased by > 30%,
  - Reduced bandwidth/latency variability
  - Near-native performances

Acknowledgement

- DOE Office of Science Advanced Scientific Computing Research award DE-SC0005050
- NSF grants CNS-0707365
- Scalable System Lab in University of New Mexico

Contact Information

Zheng Cui
Department of Computer Science
MSC01 1130
University of New Mexico
Albuquerque, 87131

Email: cuizheng@cs.unm.edu
zcui293@gmail.com

http://cs.unm.edu/~cuizheng

Questions?